Integrated Retail Analytics for Store Optimization and Demand Forecasting

1. Introduction

1.1 Project Background

Retail organizations generate large volumes of sales data influenced by promotions, seasonal trends, and external economic factors. Extracting meaningful insights from this data is critical for improving store performance, optimizing inventory, and enhancing customer experience.

This project focuses on building an integrated retail analytics solution using data analysis, machine learning, and time-series forecasting techniques.

1.2 Project Objectives

The key objectives of this project are:

- Identify anomalies in weekly sales data
- Understand seasonal and holiday-driven sales patterns
- Segment stores based on sales behavior and economic sensitivity
- Infer market basket associations at the department level
- Forecast future demand using statistical and machine learning models
- Analyze the impact of external economic factors on sales
- Propose actionable inventory and personalization strategies

2. Dataset Description

2.1 Overview of Datasets

The analysis is performed using the **Walmart Retail Dataset**, consisting of three data files:

1. Sales Dataset

- Store ID
- Department ID
- Date

- Weekly Sales (Target Variable)
- Holiday Indicator

2. Features Dataset

- Temperature
- Fuel Price
- Consumer Price Index (CPI)
- Unemployment Rate
- MarkDown1 to MarkDown5
- Holiday Indicator

3. Stores Dataset

- Store ID
- Store Type (A, B, C)
- Store Size

2.2 Data Characteristics

- Sales data contains over 400,000 weekly records
- Promotional MarkDown features contain missing values
- Economic indicators show gradual temporal variation
- No individual customer transaction data is available

3. Data Integration and Preprocessing

3.1 Data Merging

The datasets were merged using Store and Date as keys to create a unified analytical dataset. A left join strategy was applied to preserve all sales records.

3.2 Missing Value Treatment

- MarkDown features were filled with zero, representing absence of promotions
- CPI and Unemployment values were forward-filled due to their slow temporal change
- Lag and rolling features were filled with zero where historical values were unavailable

This approach ensures data consistency while preserving business meaning.

3.3 Feature Engineering

New features were created to improve model performance:

- Year, Month, and Week indicators
- Sales normalized by store size
- Lagged sales features (1-week and 4-week lag)
- Rolling averages (4-week and 12-week windows)

4. Exploratory Data Analysis (EDA)

4.1 Sales Distribution Analysis

Weekly sales exhibit a right-skewed distribution with extreme values, indicating the presence of anomalies and seasonal spikes.

4.2 Time-Based Trends

Sales show strong yearly seasonality with recurring peaks, particularly during holiday periods.

4.3 Holiday Impact

Average weekly sales during holiday weeks are significantly higher compared to non-holiday weeks, confirming the importance of holiday-driven demand.

4.4 Store Type Analysis

- Type A stores generate the highest sales volume

- Type C stores show lower and more volatile demand
- Store size plays a major role in overall sales performance

5. Anomaly Detection

5.1 Statistical Anomaly Detection

Z-score based detection was used to identify extreme deviations in weekly sales. Sales values with absolute Z-scores greater than 3 were flagged as anomalies.

5.2 Machine Learning-Based Anomaly Detection

Isolation Forest was applied to detect non-linear and complex anomalies. Approximately 1% of the data was identified as anomalous, aligning with expected retail variability.

5.3 Anomaly Handling Strategy

Rather than removing anomalies, they were retained and flagged, as many anomalies correspond to genuine business events such as promotions and holidays.

6. Time-Series Analysis and Seasonality

6.1 Trend and Seasonality Decomposition

Seasonal decomposition revealed:

- A stable long-term trend
- Strong yearly seasonality
- Residual fluctuations linked to promotions and economic changes

6.2 Store-Level Temporal Behavior

Individual stores exhibit varying levels of volatility and seasonal sensitivity, supporting the need for store-specific forecasting strategies.

7. Store Segmentation Analysis

7.1 Segmentation Approach

Stores were segmented using KMeans clustering based on:

- Average weekly sales
- Sales per store size
- Promotional activity
- Economic indicators

7.2 Segmentation Evaluation

The silhouette score (~0.24) indicates moderate cluster separation, which is expected in real-world retail datasets with overlapping behaviors.

7.3 Cluster Interpretation

- **High-Volume Stable Stores**: Consistent demand and strong performance
- **Promotion-Driven Stores**: High sensitivity to markdowns
- **Low-Volume / Volatile Stores**: Demand influenced by economic conditions

8. Market Basket Analysis (Inferred)

8.1 Methodology

Due to lack of transaction-level data, market basket analysis was inferred using department-level sales co-movement over time.

8.2 Insights

- Strongly correlated departments present cross-selling opportunities
- Bundled promotions and optimized store layouts can leverage these associations

9. Demand Forecasting

9.1 Forecasting Models Used

- Baseline Moving Average
- ARIMA (Time-Series Model)
- Random Forest Regressor (Machine Learning Model)

9.2 Model Performance

The Random Forest model achieved the lowest RMSE, outperforming both baseline and ARIMA models by capturing non-linear relationships and external factor impacts.

9.3 Forecasting Insights

Machine learning-based forecasting provides superior accuracy and is better suited for real-world retail demand planning.

10. Impact of External Factors

10.1 Correlation Analysis

- Fuel prices show mild negative correlation with sales
- CPI reflects inflationary influence on consumer spending
- Unemployment negatively impacts demand in sensitive regions

10.2 Feature Importance

Lagged sales and rolling averages are the most influential predictors, followed by economic indicators.

11. Personalization and Inventory Strategies

11.1 Segment-Based Inventory Planning

- High-volume stores require higher base inventory

- Promotion-driven stores benefit from targeted markdowns
- Volatile stores require conservative stocking

11.2 Marketing Strategies

- Personalized promotions based on store segment
- Cross-selling using department associations
- Holiday-focused campaigns for peak demand

12. Real-World Challenges and Limitations

- Absence of customer-level transaction data
- Lagging nature of economic indicators
- Impact of unforeseen events such as supply disruptions

13. Conclusion

This project demonstrates a comprehensive retail analytics framework integrating anomaly detection, segmentation, market basket analysis, and demand forecasting. By incorporating external economic factors and machine learning models, the solution provides actionable insights for inventory optimization, personalized marketing, and strategic retail decision-making.

14. Submission Notes

- All analysis was performed individually
- Code is available in the linked GitHub repository
- Supporting visuals and explanations are included